EmoDB Speech Classification

First Model

Format: As we deal with wav files and need to tackle a speech recognition challenge, mfcc format was chosen as the sate of the art format for speech recognition

Model: the architecture was inspired by this article which achieved a validation accuracy of 68% on. It consists of a 4-layers CNN treating the mfcc data as a single-channeled image.

Data Cleaning: the audio files lengths were zero padded/trimmed to reach the average signal size

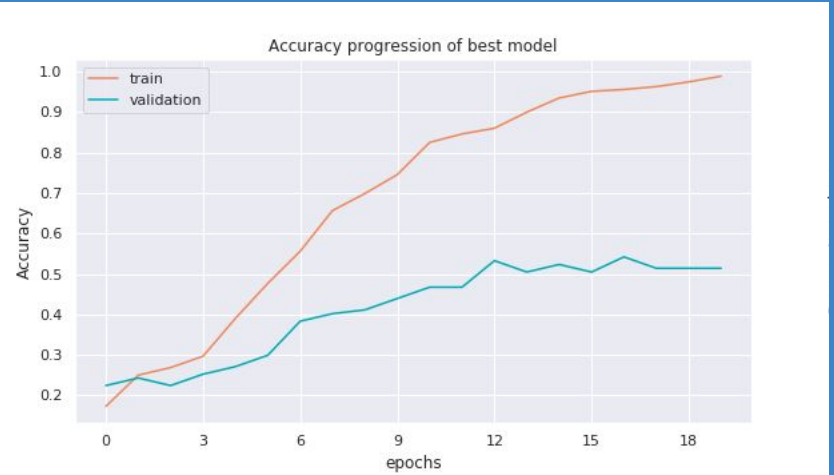
Hyper-parameters: the hyper parameters were tuned by GridSearch using the skorch library

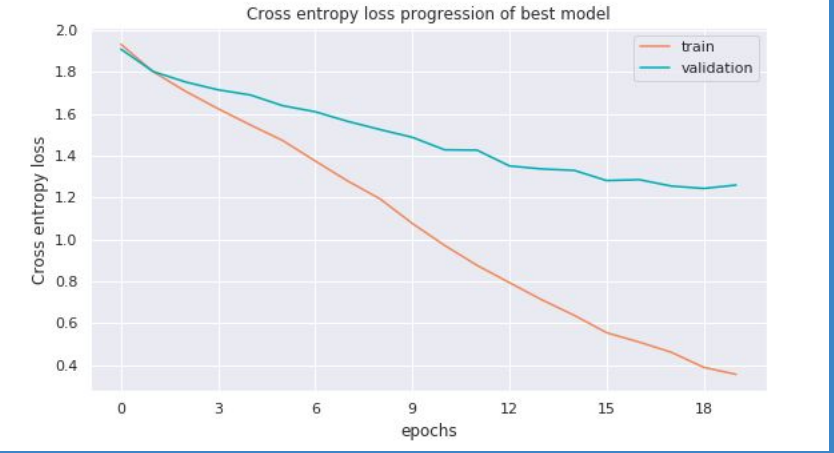
First Result

As we can see, the validation accuracy reaches a ceiling around 50%, indicating an overfitting issue. The network seems to have enough representational power still, but need some type of regularization/data augmentation.

In order to tackle this issue, 3 feature augmentation techniques were implemented, along with Dropout regularization

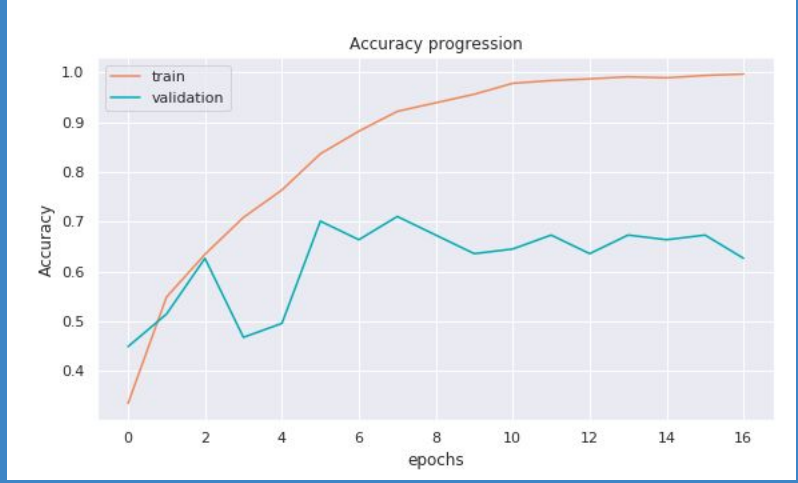
- Temporal Shift:
- Gaussian Noise addition:
- Pitch tuning





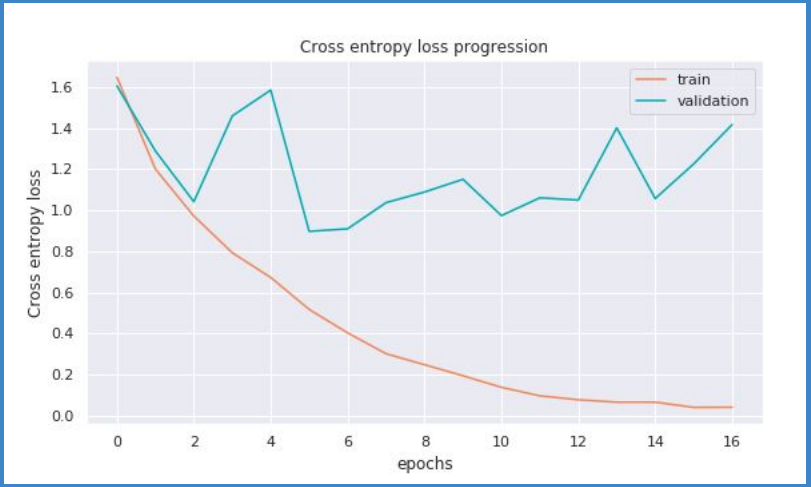
Regularized Model

The regularization seems to have allowed the validation to increase of ~10%
(3 Train/Val CV, 15 epochs)



Final Model:

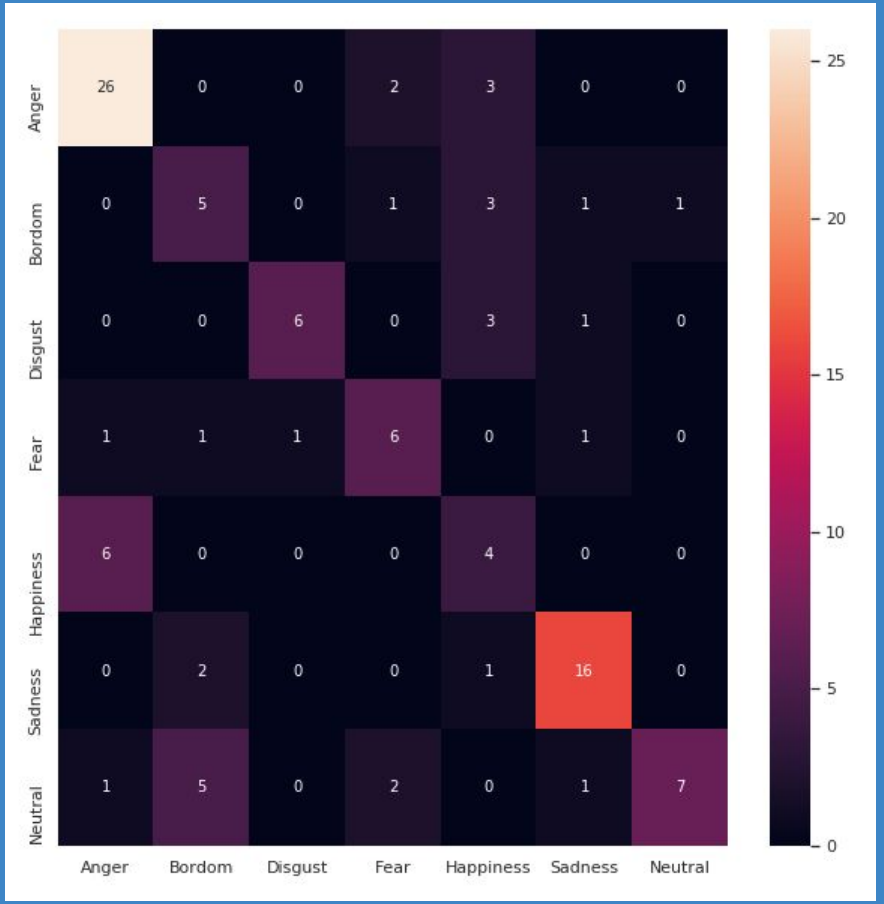
- train accuracy: 98+- 2.3%
- val accuracy: 58+- 3.7%



Misclassification Analysis

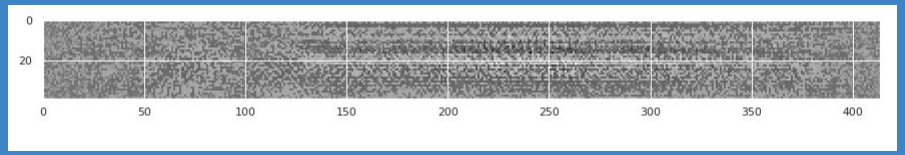
It seems like the main responsible features for bad validation accuracies are Bordom interpreted as Neutral and Happiness as Anger.

This does make sense, as the 2 formers emotions are pretty quietly pronounced whereas the latter are more energetic, suggesting that the network potentially focuses on the exclamation rate to classify the samples.



Max Activation Layer

In order to gain insights on the type of features the network focuses on to do its prediction, we use a gradient descent algorithm on the pixels of the input image to maximize the ouput of the various layer. The goal is to recover a mfcc which would show a particular pattern.



Unfortunately, due to relatively poor validation accuracy, the generated image does not reveal any significant feature. However this could happen with better hyper-parameter tuning.

What’s Next?

- Ideally, LSTM and RNN more generally would be suitable architectures to explore due to the sequential nature of the data.
- Some other data augmentation could as well increase the validation accuracy, such as random cropping/acceleration deceleration